SAND98-2211J

, , 1–20 ()

Knowledge Mining With VxInsight: Discovery Through Exploration

GEORGE S. DAVIDSON, BRUCE HENDRICKSON, DAVID K. JOHNSON, CHARLES E. MEYERS AND BRIAN N. WYLIE

{gsdavid,bahendr,dkjohns,cemeyer,bnwylie}@sandia.gov
Sandia National Laboratories, Albuquerque, NM

Abstract. The explosive growth in the availability of information is overwhelming traditional information management systems. Although individual pieces of information have become easy to find, the larger context in which they exist has become harder to track. These contextual questions are ideally suited to visualization since the human visual system is remarkably adept at interpreting large quantities of information, and at detecting patterns and anomalies. The challenge is to present the information in a manner that maximally leverages our visual skills. This paper discusses a set of properties that such a presentation should have, and describes the design and functionality of VxInsight, a visualization tool built to these principles.

Keywords: information visualization, information retrieval, graphical user interface, browsing

1. Introduction

For most of history, mankind has suffered from a shortage of information. Now, in just the infancy of the electronic age, we have begun to suffer from information excess. Data overload is bound to get worse. An urgent need exists for tools to help manage, extract information, and accumulate knowledge from increasingly large collections of data, now being accumulated and made available from data warehouses. We have rather limited experience with such problems and existing tools are often hard to use, incomplete and generally inadequate.

As we have begun to use computers to manage information, it probably can't be helped that our first thoughts are to automate the way we have done things in the past. This is the basis for online card catalogs, computerized abstract and title index services, and other tools that greatly speed up the steps we have traditionally used to work with libraries and other large databases. However, as Bar and Borrus have noted, "one reason Information Technology investments have not translated into higher productivity is that they have primarily served to automate existing tasks. They often automate inefficient ways of doing things" (Bar & Borrus, 1993). While such approaches are of undeniable value, they have proved insufficient to solve the problem of information overload. New insights and techniques are required. Fortunately, they can be enabled by the same technological advances that created the glut of data.

One of the key shortcomings of information management systems is in the nature of the user interface. Interpreting textual information is a sequential and slow process, which imposes a limit on the speed with which information can be assimi-

lated and processed. Text is excellent for conveying detailed information, but it is poorly suited for conveying relationships or trends or for getting an overview of a set of data. For these tasks, graphical displays are more effective and have become essential elements of scientific research (Tufte, 1992).

Graphical displays can be effective at conveying information because of the remarkable capabilities of the human visual cortex. Compared to the eons of evolutionary progress in vision, our talents for interpreting text and speech are recent and primitive. The visual system is incredibly effective at identifying trends and patterns, and detecting anomalies in large sets of data. These innate skills have the potential to significantly alleviate the problem of information overload. For this potential to become a reality, graphical interfaces must be developed for all types of abstract information. These interfaces must be carefully designed to leverage the inherent human strengths in our visual systems. However they must also be modest, only exploiting visualization where appropriate, because many kinds of information are best conveyed through other channels. We have attempted to build such a tool, initially to address specific questions, but which we have come to realize is very broadly applicable.

Our VxInsight tool is a graphical interface to large databases. It presents a visual representation of the data elements, but importantly, it shows them in groups that share certain similarities. Related database objects are located on a 2-D plane with a proximity based on a measure of their similarities. The resulting aggregation of elements into groups is the *context* within which an analyst begins to understand *the implicit structure of the data, rather than just asking about its content.* Context is a critical component of knowledge, which is easily lost in a deluge of raw data. Fortunately, we can represent detail while simultaneously displaying context in a graphical manner.

VxInsight overlays the 2-D plane with a 3-D virtual landscape that looks like a mountain range. The height of a mountain is proportional to the density of objects beneath it. This 3-D environment is readily explored because there is only a small cognitive step between seeing the presented images and then exercising our innate human expertise in navigating through real landscapes.

VxInsight allows one to see the big picture, to zoom in and see the details right down to the network of relationships among the data elements. If the database represents data that has accumulated through time, VxInsight also allows the analyst to see the terrain representations evolve temporally as the data set increases through its collection lifetime. These are intriguing, and compelling views of the data. However, the real power of VxInsight becomes apparent when the analyst begins to ask questions using the Standard Query Language (SQL) interface to the database. Typically, the analyst will compose a Boolean query that is passed back to the database engine. VxInsight uses the response from the database engine to visually mark the data elements that match the query. It does so by coloring spots (on the 3-D terrain representation) above the matched data elements. The results of each query are shown with different colors. This places query results in the context of the overall information terrain, and in relation to other queries, thereby presenting a powerful way to answer questions. This representation shows

not only the matching contents of the database, but also says important things about the intersection between the analyst's question and the implicit structure of the database. It is this participatory interaction with the data that gives VxInsight particular power.

In the next section we present a more detailed overview of VxInsight and discuss its relationship to previous work. In Section 3 we discuss placing the data objects into the 2-D plane and present several algorithms for this problem. In Section 4 we describe how an analyst uses the system. We follow with a discussion of current and future applications in Section 5, and conclusions in Section 6.

2. An Overview of VxInsight

VxInsight provides a visual mechanism for browsing, exploring and retrieving information from a database. The graphical display conveys information about the relationship between objects in several ways and on multiple scales. In this way, individual objects are always observed within a larger context.

For example, consider a database consisting of a set of scientific papers. Imagine that the papers have been organized in a two dimensional geometry so that related papers are located close to each other. Now construct a landscape where the altitude reflects the local density of papers. Papers on physics will form a mountain range, and a different range will stand over the biological papers. In between will be research reports from biophysics and other bridging disciplines. Now, imagine exploring these mountains. If we zoom in closer, the physics mountains will resolve into a set of sub-disciplines. Eventually, by zooming in far enough, the individual papers become visible. By pointing and clicking you can learn more about papers of interest or retrieve their full text.

Although physical proximity conveys a great deal of information about the relationship between documents, you can also see which papers cite which others, by drawing lines between the citing and cited papers. For even more information, you can choose to highlight papers by a particular researcher or a particular institution, or show the accumulation of papers through time, watching some disciplines explode and other stagnate.

VxInsight is a general purpose tool, which enables this kind of interaction with abstract databases. It allows users to interactively browse, explore and retrieve information from the database in an intuitive (by design) way.

Several basic principles guided the design of the tool.

1. The human visual system excels at identifying patterns, trends and anomalies.

2. A human can better interpret data with presentations built on familiar metaphors.

3. A useful tool allows easy and intuitive navigation through details and higher level views.

4. Visually examining large datasets is easier when it is simple to adjust the level of detail in the scene, for instance, by zooming into and out of the representation.

5. Graphical displays should have compact representations to facilitate interactive frame rates and access over a network. However, simplicity must not prevent the tool from being applicable to large datasets.

In VxInsight, we use geometric proximity as a metaphor showing the relatedness of two objects, and present the result as 3-D landscape. This is a very intuitive mechanism, but it has significant limitations. Databases typically encode many different kinds of relationships, so proximity in a low-dimensional space is insufficient to fully capture all of the complexity in the relationships. Although placing the objects in a higher-dimensional geometric space would allow proximity to convey more information, users have great difficulty interpreting and navigating higher dimensional spaces.

We feel this the ease of interaction in 3-D compensates for this compression of information. Nevertheless, the process used to determine where objects will be placed is critical to the quality of information preserved for presentation. We discuss this process in Section 3.

2.1. Related Work

A number of research efforts and commercial applications have applied visualization techniques to databases. Many of these efforts have been domain-specific, but some have been more general purpose. We will not survey this rapidly expanding field, but we will review the projects that overlap with VxInsight in philosophy or in approach.

For commercial products that strive to extract useful information from large collections of data, see, for instance the comparison articles (Edelstein, 1997, Ginchereau, et al. 1997). None of the principle commercial products make intensive use of high-performance graphical displays. The more interesting research in visualizing abstract information hasn't yet penetrated the commercial market.

In (Lin, 1997), four types of visual displays for information retrieval are described. These are hierarchical, network, scatter and map displays. Three of these types are contained within VxInsight. As will be detailed in Section 4, links between objects can be displayed as appropriate, which constitute a network display. The zoom capability within VxInsight produces a dynamic hierarchical view of the data. But the principle view within VxInsight is closely related to scatter displays. In traditional scatter displays, data objects are represented as points in a two-dimensional plane. But this representation becomes intolerably crowded when there are a large number of objects. Our solution to this problem is to display the density of objects as a landscape, only displaying individual objects when requested or when their quantity is modest. A number of other projects have addressed this problem in different ways.

For example, (Cook & Buja, 1997) suggests mapping objects to points in 3-D, but only visualizing 2-D projections. The analyst can then control the projection to identify the features of greatest interest, or to find the most informative display. The flexibility of this approach is attractive, and more information can be encoded

in the proximity relationships in 3-D. But if the dataset is large, the fundamental problem of an overly crowded display remains.

The Visual Insights project from Lucent Technologies (Visual Insights) has more in common with our approach. They use a 2-D scatter or network display, but they enable the analyst to zoom into regions of interest. This substantially resolves the overcrowding issue. However, without any mechanism for identifying the different screen regions, navigation is difficult; an analyst has trouble knowing where to zoom. But as well will discuss in Section 4, dynamic peak labels in our landscape provide important navigational guidance.

Several tools have been developed that extend scatter displays to three dimensions by placing a spike on top of each object. Two examples are the MineSet software from SGI (Rathjens, 1996) and the SDM package from Carnegie Mellon (Chuah, et al. 1995). The height of the spike can be used to convey information about the objects, and navigation in 3-D enables exploration of the data. As with the Visual Insights tool, no capacity for automatically identifying areas of the display are included.

Other projects, developed independently and concurrently with VxInsight, have taken the approach we advocate: forming a density-based landscape. In the work of (Girardin, 1996), a landscape representing the World Wide Web is constructed in much the same way as we would do with VxInsight. First, the objects are positioned in the plane using a self-organizing map approach. In contrast to our approach, a self-organizing map restricts object locations to lattice points. Next, density is used to specify an altitude for that region of the landscape. However, zooming and peak labeling techniques are absent in this work. A closely related effort is the WEBSOM project (Honkela, et al. 1996, Honkela, et al. 1998). Self-organizing maps are again used to position objects, but instead of a 3-D landscape, a 2-D display is produced in which color represents density. However, peak labels are automatically generated, and some very limited navigational and retrieval capabilities are provided.

The work most similar to our own is the SPIRE project (Wise, et al. 1995), which originated at the Pacific Northwest Laboratory and is now being commercialized by ThemeMedia. SPIRE has two visualization approaches. The first is a landscape overlaying a scatter display with peak labels, very similar to our own. The second is the scatter display itself. However, unlike VxInsight, these are separate tools, and are not unified into a single presentation. Also, SPIRE does not allow navigation or zooming, which we have found to be very useful in exploring and understanding the implicit hierarchical structure of data.

3. Geometric Placement of the Objects

A guiding motivation behind VxInsight is the use of geometric proximity in the graphical user interface to capture relationships between objects. Thus, a critical step is the mapping of objects to geometric coordinates in such a way that related documents are kept close together. The ability to extract useful information from the tool depends upon the quality of this *ordination*. The properties of a good ordination depend upon the nature of the data and the questions being asked, so it is

our conviction that no single approach is a panacea. With domain-specific expertise, an analyst may be able to provide a better ordination than any general purpose tool. For this reason, we allow the analyst to choose among several ordination algorithms, or to bypass our algorithms and provide coordinates directly.

3.1. The Similarity Function

Input to the ordination algorithms in VxInsight consists of a *similarity* function $s : O \times O \rightarrow \mathbb{R}$, which maps object pairs to non-negative real numbers. Larger values imply that the two objects are more similar and hence should be located closer together. For large datasets, we expect most of the similarity values to be zero, so the data representing the similarity relationship will be sparse.

The details of the similarity function depend upon the application. Some simple examples of the data that could be used for similarity generation include the following.

1. Common keywords in documents.
2. Identical vocabulary within documents.
3. Citation links between scientific papers or patents.
4. Direct links in web documents.
5. Financial transaction links between corporations.
6. Membership in common organizations among individuals.

More generally, because any relational database includes information that couples objects, the different relational fields can be summed or combined in more complex ways to generate similarity values. Of course, more sophisticated methodologies, like latent semantic indexing, (Deerwester, et al. 1990) could also be used. Obviously, the analyst should select a function that is appropriate for the kinds of questions being addressed.

3.2. Classes of Ordination Algorithms

Given a similarity function, the goal of the ordination process is to place the objects in a geometric space so that similar items are close together. As we discussed above, for VxInsight we want a two-dimensional ordination, but the problem can be phrased in any dimension. Various approaches have been proposed for this and closely related problems; see, for example, the literature on self-organizing maps, eg. (Kohonen, 1997). Much of the prior work can be grouped into one of the three categories we discuss below. However, none of these general approaches is a perfect match for our needs.

In **graph drawing**, the objective is to sketch a graph in the plane in a visually pleasing manner. There is an annual conference devoted to this topic, and numerous

algorithms and applications can be found in the proceedings, eg. (Goos, et al. 1997). Possible objectives include non-overlapping vertices, few crossing edges and short edge lengths. The similarity data in our ordination problem can be interpreted as a weighted graph, in which objects are vertices and non-zero similarities are edges with similarity values as weights. Graph drawing techniques can now be applied to this graph.

However, the visual display of the graph of similarities is not of paramount concern. As will become clearer when we describe the visualization process in Section 4, the principle visual paradigm in VxInsight is the implicit clustering generated by the ordination process. In fact, the graph of similarity values might never be displayed. For this reason, graph drawing is not the right paradigm for our ordination process.

Another possible approach would be to use one of the many **clustering** algorithms (see, for example, the survey paper (Willett, 1988)). Specifically, large clusters could be identified and relegated to different portions of the geometric space. The subclusters within them could be identified and ordinated recursively.

Although appealing, this approach goes against the guiding philosophy of VxInsight. We feel that the human visual system is better than any algorithm at identifying patterns and trends. We don't want the clusters to be imposed by an ordination algorithm, but rather identified by the analyst as an implicit product of the ordination algorithm. To put it another way, an ordination algorithm that is based upon the detection of clusters may miss other features that are equally important to the analyst.

A third possible approach to the ordination problem is to use techniques from **multidimensional scaling** (MDS) (see, eg. (Borg & Groenen, 1997)). The fundamental problem in MDS is to find a low-dimensional ordination for a set of objects that is consistent with some pairwise distance information.

Although this problem is closely related to our own, there are two reasons why we chose not to use techniques from this field for VxInsight. First, we expect our similarity data to be of low fidelity – mere hints about what is desired. Thus, we don't have any good input concerning distances to input to MDS. Second, the techniques used in MDS are computationally intensive, and we wish to be able to handle very large datasets, ideally in real time.

3.3. The Ordination algorithms in VxInsight

VxInsight currently contains two ordination algorithms with complimentary strengths and weaknesses. The first involves eigenvectors of a Laplacian matrix and is closely related to multidimensional scaling. The second is a particle algorithm, similar to molecular dynamics, in which objects move about under attractive and repulsive forces. The eigenvector approach has the attractive property that it finds the global minimizer of a reasonable objective function. However, the ordinations it produces tend to be too tightly clustered, and so not ideal for interactive visualization. The particle method produces more visually appealing ordinations, but it tends to get

stuck in local minima, and so does not produce the best possible ordination. In our experience, the combination of these two algorithms is better than either alone.

3.3.1. Laplacian Eigenvectors One way to force similar objects to be close together is to minimize an appropriate penalty function. Many such functions are possible. However, many, also, lead to intractable computational problems. The particular approach described here reduces to a symmetric eigenvalue problem; a problem for which good software and algorithms are available.

Consider the following penalty function on the n objects

$$\text{Cost} = \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij} d_{ij}^2, \tag{1}$$

where s_{ij} is the non-negative similarity between objects i and j and d_{ij} is the geometric distance between them. Minimizing this function will encourage highly similar objects to be close together. However, merely minimizing Eq. 1 leads to a poorly phrased mathematical problem. Several constraints need to be added. First, note that translations of the full set of objects doesn't alter the cost. We can resolve this by adding constraints of the form $\sum_i x_i = 0$ and $\sum_i y_i = 0$, where x_i and y_i are the x and y coordinates for object i.

With these translational constraints, the minimum cost is trivially obtained by placing all the objects at the origin. This uninteresting solution can be avoiding by adding the constraints $\sum_i x_i^2 = 1$ and $\sum_i y_i^2 = 1$. The solution to the resulting minimization problem will place all the objects along the main diagonal $x = y$. This can be avoided by adding a constraint of the form $\sum_i x_i y_i = 0$. Putting these constraints together, we have the following well posed minimization problem.

$$\text{Minimize} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}\{(x_i - x_j)^2 + (y_i - y_j)^2\} \tag{2}$$

Subject to :

$$\sum_{i=1}^{n} x_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} y_i = 0$$

$$\sum_{i=1}^{n} x_i^2 = 1 \quad \text{and} \quad \sum_{i=1}^{n} y_i^2 = 1$$

$$\sum_{i=1}^{n} x_i y_i = 0$$

As discussed in (Hendrickson & Leland, 1995), the solution to this problem involves the *Laplacian matrix* of the similarity function, L. Values in this matrix are defined as follows.

$$L(i,j) = \begin{cases} -s_{ij} & \text{if } i \neq j \\ \sum_{j=1}^{n} s_{ij} & \text{if } i = j \end{cases} \tag{3}$$

Denote the n eigenvectors of L by u_i, where the corresponding eigenvalues are ordered from smallest to largest. Then the solution to Eq. 2 can be shown to be $x = u_2$ and $y = u_3$ (see, for example, (Hendrickson & Leland, 1995)).

The eigenvector problem has been well studied and several good software tools for this kind of calculation exist. We have chosen to use the ARPACK code by (Lehoucq, et al. 1998) due to its ability to limit the memory requirements. Using ARPACK with Tchebycheff polynomial preconditioning, we were able to ordinate a database with 2.4 million objects in less than a day on a high-end workstation.

Although Laplacian eigenvectors are computable in reasonable time and achieve the global minimizer of a plausible objective function, they have several shortcomings for our purposes. First, there are degenerate eigenvectors unless the graph with nonzero similarity values is *biconnected.* That is, the graph can't be divided into unconnected pieces by removing a single edge. In practice, this shortcoming can be handled in one of two ways. Either the biconnected components of the graph are handled up front (for a linear time algorithm, see, for example, (Aho, et al. 1974)), or edges can be added to make the graph biconnected. Note that if a graph is connected, adding edges to shortcut all length-two paths will make it biconnected.

The second, and more serious, problem associated with Laplacian eigenvectors concerns the type of ordination they produce. In our experience, the objects tend to be highly clustered – so much so that navigation and interpretation are impaired. So for large datasets, we have found this approach to be unsuitable without some modification.

3.3.2. Particle Methods for Ordination In the second ordination algorithm currently implemented as part of VxInsight, objects are moved about under the influence of attractive and repulsive forces. Specifically, each object is attracted towards the objects to which it is similar. The strength of this force is proportional to the similarity value. Simultaneously, a repulsive force exists between each pair of objects that are too close together. There are many possible variations of force laws within this basic model.

This *particle* approach has several attractive features. It generally produces ordinations that are attractive for interpretation and navigation. It is also good for incremental ordinations, where a few new objects are being added to an existing landscape.

However, this method also has limitations. Unlike the Laplacian eigenvector approach, the particle method is sensitive to initial starting conditions and can generate a different answer with only small differences in initial data. Further, it can get trapped in a local minimum, and never find the globally optimal solution. Also, particle methods are very slow to compute, although tricks from the computational physics community can reduce computation time (Allen & Tildesley 1987).

Despite their individual draw backs, together the Laplacian eigenvector algorithm and particle method complement each other. We have had success by first using eigenvectors and then refining the ordination using the particle method. In practice, the particle method spreads the tightly clumped eigenvector ordination very nicely. Pairing these complementary methods produces an acceptable ordination, which

requires only modest computing resources. Also, by initializing the particle method with the globally good ordination from the eigenvector approach, the problem of multiple minima is reduced.

4. Using VxInsight

VxInsight is more than the user interface, but the features of the user interface are what make it truly powerful. However, behind the scenes, a subset of a larger database may have been selected, considerable processing may have been required to compute the similarity measures between elements, and of course the ordination had to be computed. All of this work must have happened before the analyst brings up the interactive part of VxInsight. Then, once the preliminaries are completed, the interaction with the user begins; real value starts being added by the analyst's interactions.

In this section we will describe some of the physical architecture behind the system, show the architecture of the client/server interface to the database engine, discuss how the data is pre-processed and then begin describing what the user finally sees. That description will address what the presented environment looks like, how it can be personalized with various customizations, and how one operates the features of the graphical user interface. We will explain our concepts for presenting multiple levels of detail. We will also discus how SQL is used to directly access the database for user specific queries, and how the results of these queries are graphically displayed. We will explain how VxInsight automatically configures itself, changing the names of available fields to match the current database, of which several can be available at any given time. Finally, this section will close with a description of a typical analysis session.

4.1. The Physical View (Network, Workstations, Data Storage)

Figure 1 shows an analyst connected to multiple databases over the company intranet. It also depicts the relationship between the database server, and the VxInsight server. This figure suggests that most jobs will require access to multiple databases, and that switching between them or combining them should be easily accomplished. These databases are likely to be accessed using web technologies, but they are almost surely protected behind the company firewalls, because they are valuable assets.

Notice that the database server is shown residing on the same machine as the VxInsight server, with socket connections between them. This is, however, not necessary. In fact the database server could be widely separated from the VxInsight server, the socket connections will make this physical separation transparent to the analyst using the system.

The database server is likely to be one of the major commercial database systems. Which one is not critical, as long as it supports SQL; we have used Oracle software for our applications. That server interfaces to the VxInsight server, which provides data to the client software at the analyst's workstation. It serves multiple analysts

at once. It also accepts the SQL from the clients and makes the actual database query, which is slightly system dependent.

4.2. Controlling the Display

A typical study will only require access to a subset of the entire dataset, so the first preprocessing step is to pull out the required data elements, and create a more manageable database. This new database is extended to have two new tables, one for the X coordinate, and one for the Y coordinate. these coordinates are used to construct the terrain representation. By carefully controlling the resolution of the various views, the graphical rendering loads are kept below the interactive frame rate. Speed is a part of the user interface, and hard tradeoffs are required between performance and affordable computing equipment. Once the images of the landscape have been computed, the analyst begins interacting with the system, usually by navigating through the visual representation of the terrain, as depicted in Figure 2.

Different users have different visualization preferences and VxInsight allows considerable customization of the presented views. All of the visualization parameters can be changed at any time. For example, three kinds of mountain landscapes can be requested, a transparent wireframe rendering (shown in Figure 3), which lets the analyst see the density of data elements below the mountain, or solid mountains, which can either rise from a synthetic sea, like an island (as in Figure 1), or can rise up from an interior grassland region. Changing to the wireframe view is particularly useful when various other customizations are enabled.

For instance, one option enables drawing lines between the individual data elements to directly show their actual linkages (for example, the citing/cited relationships between scientific papers). These can be seen in their proper context, underneath the wireframe representation.

Another useful customization allows the heights of the mountains to be scaled linearly with data element density, or logarithmically with that density. This is particularly useful when the data has a wide range and linear scaling would cause the smallest mountains to be lost in the scale of the largest concentration of data elements, which would have the highest mountain elevation.

Still another useful customization allows the analyst to specify how peaks should be labeled. For instance, some datasets can be labeled clearly with 20 character names, others will require longer names. Also, the analyst can specify which fields should be used to create the peak names. For example, the peaks might be named using common words found in the journal titles for the elements under the given peak. Alternatively, the names could come from common words in the paper titles, or in the abstracts.

Different users are comfortable with different rates for zooming. A very high zooming speed for one person may be seen as a time saving convenience, while the same speed may make another analyst uncomfortable. This rate is easily customized to fit the preferences of each analyst.

4.3. Navigating in Landscapes

Navigation tools for visual interfaces span a huge range of capabilities; from no interactive capability to complete, virtual reality interfaces even including force feedback and sound representations of the data and auditory clues about how fast you are moving in the virtual environment. We have tried various combinations of navigation techniques. Interestingly, we found that providing too many degrees of control for navigation was harmful; users wound up be lost, flying upside down, or frustrated by the complication of the more powerful navigation paradigms. We have finally selected a small set of motions, all of which are simply controlled using a mouse. We allow the user to control the distance from the terrain, and to fly about the terrain to change the viewing angles.

We allow four very simple navigation gestures. First, moving the mouse forward moves you closer to the terrain. Second, moving it back pulls away from the terrain. Third and fourth, moving the mouse to the left rotates the viewing angle to the left, while moving it to the right rotates the view to the right.

4.4. Levels of Detail

The real power of information visualization comes with the capability to view the data on different scales. One intuitive concept of level of detail presentation involves zooming into a mountain range to see more detail. However, in contrast to zooming into an actual mountain, VxInsight allows you to see more and more detail, as you move in, not just a magnification of the same terrain. This means that as you move toward a mountain, it will slowly dissolve away to reveal the fractal grouping of data elements within the original large mountain representation. Figure 4 shows a sequence of images constructed by zooming into a landscape. In an interactive visualization, the images transition smoothly.

This selection of level of detail moves seamlessly from the outer most view of the 'big' picture down to the individual data elements and the visible linkages between them. That is a powerful capability, but alone it can be confusing and disorienting for the analyst. In addition to the detail shown in the current view, it is important to be able to see exactly where you are in relationship to the whole set of data. VxInsight has solved this issue by providing the user with a smaller map of the whole data region shown from the outer most view. The user can always glance at this small map in the corner of the screen and immediately see a mark representing the region he has currently zoomed into and is viewing in detail with the larger rendering.

This attention to levels of detail extends down to the individual data elements themselves. At the finest level of detail, small groups of individual data elements are visible. These are represented by small pyramids. They are colored to represent various auxiliary information. For instance when a query to the database is made, the results that match that query are all colored the same. Also, at this level of detail the linkages between individual data elements can be colored to represent the direction of the linkage. We color the line between a citing and a cited paper

with two colors that gradually blend together at their midpoints. This technique allows you to recognize instantly which paper is citing the other.

If zooming is one natural way to traverse levels of detail, another is to control time. VxInsight allows you to see the representation of all of the data elements in the database. However, you can also ask to see only those elements that were published within a specified time range. Using a slider on the graphical interface, this small time interval can be pulled across the entire range of dates spanned by the database, as depicted in Figure 5. As you do this, the mountains will grow and shrink reflecting the papers that were published and accumulated through that time period. This is especially useful in noting the development of trends. The images in Figure 5 show three time slices of the same data set. The colored markers denote the papers associated with particular database queries.

4.5. Querying the Database

One of the most powerful features of VxInsight is the previously mentioned ability to query a database and see the matching data element light up across the terrain with a color associated with the individual queries, as in Figure 5. For instance, at the highest level view, the analyst could use the SQL window, as shown in Figure 6, to make a database query asking to mark all the papers that were published in Japan. This might show up as blue dots directly above the associated representation of the paper (actually, just above the mountain surface, so that they are visible from outside the mountain). A second and third query might ask to mark the papers from Germany, and then from the United States. The distribution of papers from these countries are then obvious, and of course, they are still in the context of all similar papers.

When used with the time slider, visual specifications of matching queries can be used to track the ebb and flow of competitive advantage as represented in the published record.

4.6. Using Multiple Databases

Of course, most analysts will work with more than one database. They may choose to work for an extended time with one database, then turn to another one for equally long periods of work. However, it is more likely that they will want to switch between correlated databases during the same session. VxInsight provides features to support this switching seamlessly.

When a dataset is accessed a simple descriptor file is consulted. This descriptor file allows the analyst to customize the menus for a particular research focus. For example, when a dataset of web pages is being viewed, the informational display has text widgets labeled 'Title' and 'Web Address'. The peak labels can be set to either 'Title' or 'Company' and the 'Connections' menu allows us to show 'Web Links'. Now by simply opening a dataset based on financial dealings, the entire informational display and menuing options transform. The information display is now composed of text widgets reading 'Transaction Type', 'Amount', and 'Date'.

The peak label options are now 'Buyer', 'Seller','Institution', and the 'Connections' menu allows us to show either 'Country' or 'Bank' links in the data. The SQL query options will also be automatically updated. The SQL query window allows Boolean conjunctions and negations using the table names of the current database. These are immediately presented to the analyst, which is very useful when initially learning to use a new database, or when you come back to a database after an extended period of time and have forgotten the field names.

4.7. An Example Session

Now that some of the useful features in VxInsight have been introduced let's follow an analyst using the system. This example concerns a semiconductor laser technology called Vertical Cavity Surface Emitting Lasers (VCSELs). The pertinent papers would first have been extracted from the all of science database, and processed as described earlier.

The analyst starts the local VxInsight client software, which automatically connects to the VxInsight server, which is already connected to the database server. The VCSEL database is requested, and a highest-level view of that terrain is presented. The analyst asks to have the peaks labeled using words form the journal names. These clues begin to explain the visible landscape. Various concepts have their own peaks, for example fabrication technologies, applications of VCSELs, and more exotic papers with technical titles obscured with incomprehensible jargon.

In an effort to find a place to start, our analyst makes a query asking the system to mark all the papers with an author from Sandia National Labs. Suddenly, it becomes clear that fabrication technologies are important to Sandia. That's a good clue. Now the analyst zooms down into that region to reach the individual papers. He is particularly interested in finding a seminal paper, maybe one that started the whole field. Such a paper will be referenced by most of the important papers under the mountain. The analyst request that the system show the linkages by drawing lines between the citing and the cited papers. Sure enough, a few papers stand out as much more important than others.

Our analyst, wonders if most of the work is being done in the US, so he zooms back out to a higher level and makes an SQL query asking for papers from US institutions. Many of the papers are marked, but it is now obvious that the US does not have a monopoly on this technology. Now curious, the analyst asks about Japanese papers, and then perhaps other countries, until he thinks to ask if the Germans have worked in this field. Sure enough, they have a strong presence.

Even more curious, our analyst wonders if research into this area has always been spread around the world, or if the leadership has changed through time. To test this idea, he selects a time period early in the history of the technology. The early papers are almost all from Japan. Moving forward in time he notes a strong surge of American interest, though the Japanese effort never really waned. Moving to the most recent papers, he notices an almost instantaneous burst of German publications.

This is so curious that he zooms down into the more recent papers, finds a linkage where a Sandia scientist cites several new German papers. Clicking on that icon for the Sandia paper, he pulls up the text, author information, and abstract. Now rather than continue exploring the raw data, he picks up the phone and calls the Sandia researcher to ask what happened, how did the Germans suddenly become dominant in this field (some questions are best answered the old fashioned way – by asking the expert, *after you have found him*).

An actual analysis session in the field of VCSELs, though not identical to this little story, did happen and was generally in the spirit as presented. Since that time, several other studies in various fields have been undertaken, see the discussion in the applications section. It is always interesting to watch the analysts come across unexpected relationships or events, and watch them pose questions about those anomalies, and then watch as they further explore the implications of their idea searching through the data in new ways.

5. Current and Future Applications

The development of VxInsight was motivated by a simple, and universal, question: 'Where do we put the next research dollar for the most impact?' In attempting to think about an approach to this question, it became clear that an underlying prerequisite must be the development of a more robust understanding of how we got to where we are today, or as historians point out, 'those who do not understand the past are doomed to repeat it.' We wanted to avoid as many of those mistakes as possible, which meant we really had to look deeply into the historical record.

While no scientist, administrator, or policy analyst can completely *know* the world of science, we must be able to understand the technical details sufficiently to follow the evolution of the scientific disciplines if we are to direct our research funding intelligently. Our investment strategy must weigh many questions as we consider proposals to fund the proposed research of individual scientists and collaborative groups. For example, we consider questions including:

- What prior work was necessary before a breakthrough became possible?
- How do fields converge or diverge over time?
- How do new insights propagate across disciplines?
- Who are the central figures in an evolving research thread?
- In any given area, what are our competitors doing (where competitors can be individuals, institutions, or even countries)?

To begin to understand such structural issues, an obvious place to look is in the record of scientific progress documented by research publications. The evolution of ideas is contained in these papers, sometimes fully disclosed by the citation links, but sometimes only implicitly acknowledged through indirect chains of citations.

If the body of a paper contains its contributions, the paper's citations provide the recorded context – the environment in which the research evolved – and is the

author's acknowledgment of the intellectual heritage leading to the contribution presented in the paper itself. VxInsight was first conceived as a tool for examining and exploring the structure in those citation linkages. Our objective was to provide a tool with which an analyst could examine a very large set of citation information, follow the evolution of individual technologies, and, potentially, gain insight into where those technologies might be heading.

One example of the application of VxInsight to the investigation of scientific disciplines was outlined in Section 4.7. The tool has been used for several such studies. For the VCSELs study sketched above, we began with a subset of relevant papers from the Science Citation Index from the Institute for Scientific Information. VxInsight then allowed an analyst to discover who the original developers of the technology were, who then took the technology and made significant strides to improve it, and at what level other organizations and countries have continued research hoping to eventually build the world's most efficient low-power light source. Most revealing, we uncovered a systematic shift of research leadership around the world as the technology matured.

Usually the world-wide flow of technology is good, but sometimes we would rather our secrets stayed at home. For example, VxInsight was used at Sandia to study nuclear proliferation (Irwin, et al. 1997). As with the VCSELs study, scientific papers in nuclear technology and their citation links were used, to create an initial display of the structure of the nuclear technology literature. Then, text analysis tools were used to determine similarities between the scientific papers and other, public, sources discussing nuclear technologies, such as international news stories. With these new associations between the underlying science and the popular reports, those news stories were added to the landscape of scientific specialties. This coupling revealed which stories were related to sensitive information and provided a powerful, new intelligence tool. This example illustrates the flexibility of VxInsight for various types of similarity analysis.

Another group at Sandia is using VxInsight to identify potential strategic partnerships that can benefit our laboratory. They compare the relationships between external publications from other organizations with the published research of Sandia scientists. The intersection of these records, presented in a VxInsight landscape, reveals potential partners. Often the individual scientists are already aware of researchers at the indicated institutions. However, without VxInsight, managers and technology transfer experts would be unlikely to find these connections on their own. This is a powerful attribute of VxInsight – non-experts can quickly gather information that only experts, in their narrow fields, knew about or had internalized. Broader strategic visions are enabled as this information flows to managers, funding agents, and inter-company negotiators.

Scientific papers are not the only objects that have natural linkage structures. Linkages are ubiquitous, in business, in government, in social organizations; indeed most human endeavors have aspects that can be represented with this abstraction. While it is intellectually stimulating to consider the emergent order eminent in the networked abstractions of various human activities, we are more interested in

practical applications and in tools to help us make better decisions. Using VxInsight we have *seen* the implicit structures, now we are concerned with their interpretation.

What do the structures mean, and are they relatively stable or just transient phenomena? Can we analyze a structure to discover principles governing its creation and the couplings among the elements? That is, can we use these structural hints to create new and useful knowledge with practical application in decision making, and if so, then just how broadly applicable is this tool?

Accumulating experience with the database of scientific literature led us to believe that the chronicled progress and evolution of human systems, as captured in a wide variety of databases, all represent networked activities. Based on our early, but limited, experience, we also believe that many of these can be beneficially analyzed using the processes we applied to track the evolution of recent scientific research.

Some of these interesting uses of VxInsight are far from our initial interests, though we have readily recognized their importance and economic impact when visitors have shown us new applications and possible uses. Some of the suggested areas of usage include medical practice, product marketing research, counter-terrorism intelligence, monitoring the growth of national imports and exports, structuring the results of World-Wide Web searchs, and even genome mapping, where perhaps we can trace how Nature has re-used similar genes throughout the tree of life.

One obvious suggestion was to apply VxInsight to patent citations, which are very similar to citations between scientific papers. With VxInsight, a patent analyst can quickly investigate questions like: where are competitors placing their efforts, and how does their intellectual property intersect with our own? Are there areas that offer potential for patent circling? Who is citing our patents, and what types of things have they developed? Are there emerging competitors or collaborators working in areas of importance to us?

Others have pointed out that multi-dimensional transactional analysis (for example, the record of electronic funds transfer) interests certain government agencies. These databases have many fields, or tables, any combination of which might be used to study the structures within the record of transactions.

Different similarity functions can be developed, each of which weights the relative significance of particular fields as needed for individual investigations. While different similarity functions allow for different views of the data, the entire database always remains accessible via SQL queries, so that ones options to access the complete dataset are never constrained.

Of course, business, as well as government, is interested in the flow of funds. The ability to quickly understand and follow ideas and leads through a database of transactions can have huge financial implications. Currency trading, planning for international development, and detecting world-wide trading patterns all require timely information, which can be presented and studied using VxInsight. However, business and government uses represent only a small fraction of the possible applications. More broadly, we see applications in all manner of social, and management networks involving relationships between physical objects and or places. The broadest applications will surely involve abstract data, representations of concepts uncoupled from any physical manifestation. With this broader point of view,

we tried to make sure that VxInsight could be used with very general classes of networks and databases, which is now one of the its real strengths.

Anywhere there is a relationship between data elements or anywhere an abstract relationship can be defined, VxInsight can be used to explore the relationships using the natural, intuitive landscape metaphor to organize and present the implicit structures. We feel that the flexibility of our visualization paradigm and the associated ordination software will allow for a large and diverse set of applications, many of which we can not yet envision.

6. Conclusions

VxInsight is a powerful and flexible tool for interacting with large collections of abstract information. It is particularly well suited for identifying structure and patterns in large datasets, a task that is difficult for traditional knowledge management tools.

VxInsight employs an intuitive landscape metaphor, which simplifies interpretation and navigation. Geometric proximity is used to convey similarity between objects, but the precise meaning of 'similarity' is under the control of the analyst. The tool allows for zooming to explore interesting regions in greater detail, which can reveal structure on multiple scales. An SQL connection to the database allows the retrieval of detailed information about the objects. Properties of the data can also be encoded into the landscape, allowing visualizations of the distribution of objects matching a query, or the comparison of multiple queries. All this functionality runs at interactive frame rates for datasets of many tens of thousands of objects. At these rates workstations are powerful enough to allow the combination of desktop browsing, data retrieval and exploration.

The software was designed to be easily reconfigurable to work with new and different types of data. It has been used at Sandia National Laboratories in patent analysis, nuclear nonproliferation, strategic planning, World Wide Web analysis, research prioritization and other applications. In all of these settings, VxInsight has allowed the analyst to overview large collections of data, to identify interesting patterns and trends, to ask unanticipated questions, and to explore possible answers.

We believe that VxInsight is a prototype of knowledge management tools, which will heavily rely on high-performance visualization. The human visual system is extraordinarily adept at surveying large sets of information and identifying trends and exceptions. The challenge is to present abstract information in a format that allows our visual talents to excel. Based upon our experience with VxInsight, we advocate the following principles for this task.

- Use a familiar metaphor.
- Enable navigation, but keep it simple.
- Provide mechanisms for information retrieval.
- Present information at different levels of detail.

- Include multiple, simultaneous display options.
- Make it run fast on large datasets.

Acknowledgements

We are deeply indebted to Henry Small and David Pendlebury for helping define the scope of the work. We also benefited greatly from the input provided by Nancy Irwin, Helen Koller, Joyce Van Berkel and Kevin McCurley. This work was performed at Sandia National Labs which is operated for the US DOE by Lockheed Martin Corp. under contract DE-AC04-94AL85000.

References

Alfred V. Aho, John E. Hopcroft, and Jeffrey D. Ullman. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, PA, 1974.

M. P. Allen and D. J. Tildesley. *Computer simulation of liquids.* Oxford Science Publications, Oxford, 1987.

François Bar and M. Borrus. The future of networking. Technical report, University of California, BRIE Working Paper, Berkeley, CA, 1993.

Ingwer Borg and Patrick Groenen. *Multidimensional Scaling: Theory and Applications.* Springer Series in Statistics, 1997.

M.C. Chuah, S.F. Roth, J. Mattis, and J. Kolojejchick. SDM: Selective dynamic manipulation of visualizations. In *Proc. ACM Symp. User Interface Software and Technology*, pages 61–70, Pittsburgh, PA, 1995. ACM.

Dianne Cook and Andreas Buja. Manual controls for high-dimensional data projections. *J. Computational and Graphical Statistics*, 6(4), 1997.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Information Science*, 41(6):391–407, 1990.

Herb Edelstein. Mining for gold. *Information Week*, pages 52–70, April 21, 1997.

Bill Ginchereau, Julie Dunn, and Lori Mitchell. Knowledge management solutions. *Info World*, pages 116–126, November 17, 1997.

Luc Girardin. Mapping the virtual geography of the world-wide web. In *Proc. Fifth International World Wide Web Conf.* Elsevier, 1996.

G. Goos, J. Hartmanis, and J. van Leeuwen, editors. *Proceedings of Graph Drawing '97.* Springer-Verlag, 1997.

B. Hendrickson and R. Leland. An improved spectral graph partitioning algorithm for mapping parallel computations. *SIAM J. Sci. Comput.*, 16(2):452–469, 1995.

T. Honkela, S. Kaski, T. Kohonen, and K. Lagus. Self-organizing maps of very large document collections: Justification for the WEBSOM method. In I. Balderjahn, R. Mathar, and M. Schader, editors, *Classification, Data Analysis, and Data Highways*, pages 245–252. Springer, Berlin, 1998. See http://websom.hut.fi/websom/.

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki Univ. Tech. Laboratory of Computer and Information Science, 1996. See http://websom.hut.fi/websom/.

Nancy H. Irwin, Joyce van Berkel, David K. Johnson, and Brian N. Wylie. Navigating nuclear science: Enhancing analysis through visualization. Technical Report SAND97-2218, Sandia National Labs, Albuquerque, NM, 1997.

Tuevo Kohonen. *Self-Organizing Maps, Second Extended Edition.* Springer Series in Information Sciences, Vol. 30, Berlin, Heidelberg, New York, 1997.

R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK User's Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods.* SIAM, Phildelphia, PA, 1998.

Xia Lin. Map displays for information retrieval. *J. American Soc. Information Sci.*, 48(1):40–54, 1997.

Dieter Rathjens, Danny Galgani, Cindy Kleinfeld, and Christina Cary. MineSet user's guide. Technical Report 007–3214–002, Silicon Graphics, Inc., Mountain View, CA, 1996.

Ben Schneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. Visual Languages '96*. IEEE, 1996.

Eric J. Stollnitz, Tony D. Derose, and David H. Salesin. *Wavelets for Computer Graphics.* Morgan Kaufmann, 1996.

Edward R. Tufte. *The visual display of quantitative information.* Graphics Press, 1992.

Visual insights. See http://www.visualinsights.com/.

Peter Willett. Recent trends in hierarchic document clustering: a critical review. *Information Processing & Management*, 24(5):577–597, 1988.

James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proc. 1995 IEEE Symp. Information Visualization*, pages 51–58. IEEE, 1995. See http://www.thememedia.com/.

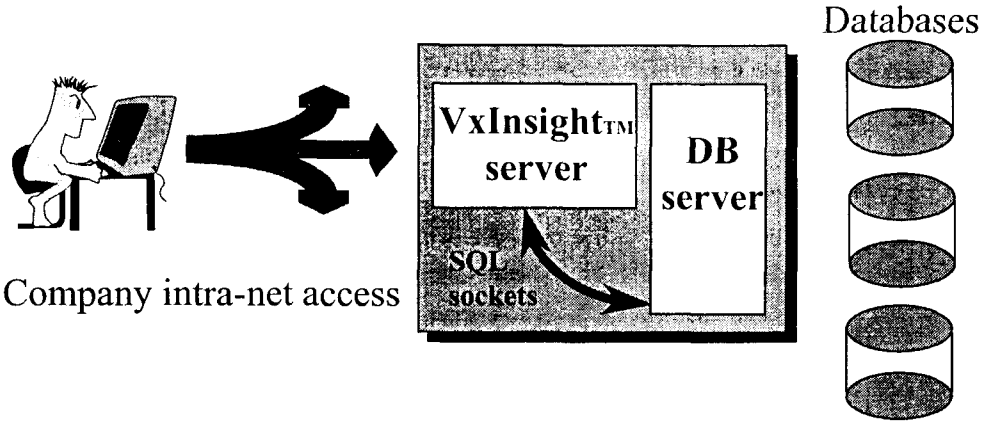

Figure 1. A sketch of the system architecture with desktop access to multiple databases via a company intranet. The VxInsight server and the database server can reside on the same machine, or can be widely separated computers.

Figure 2: The mountain terrain metaphor provides an intuitive exploration environment.

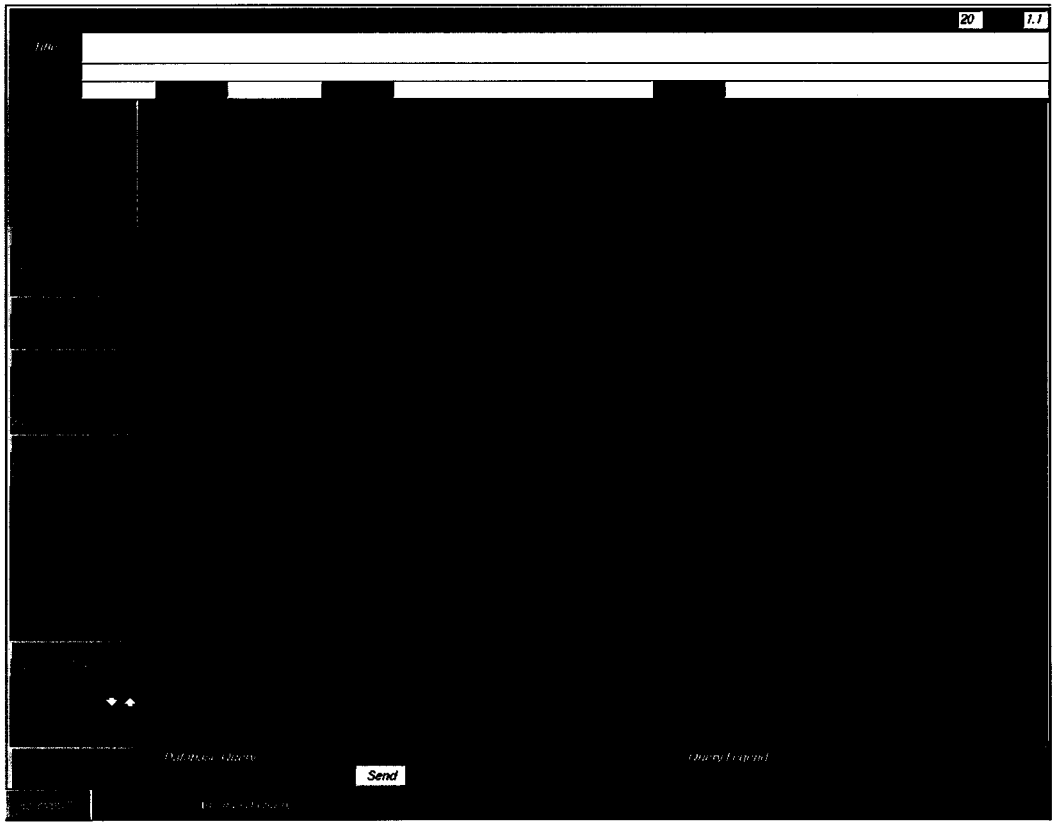

Figure 3: Dynamic terrain generation: Height is based on the density of data objects (shown as gray dots).

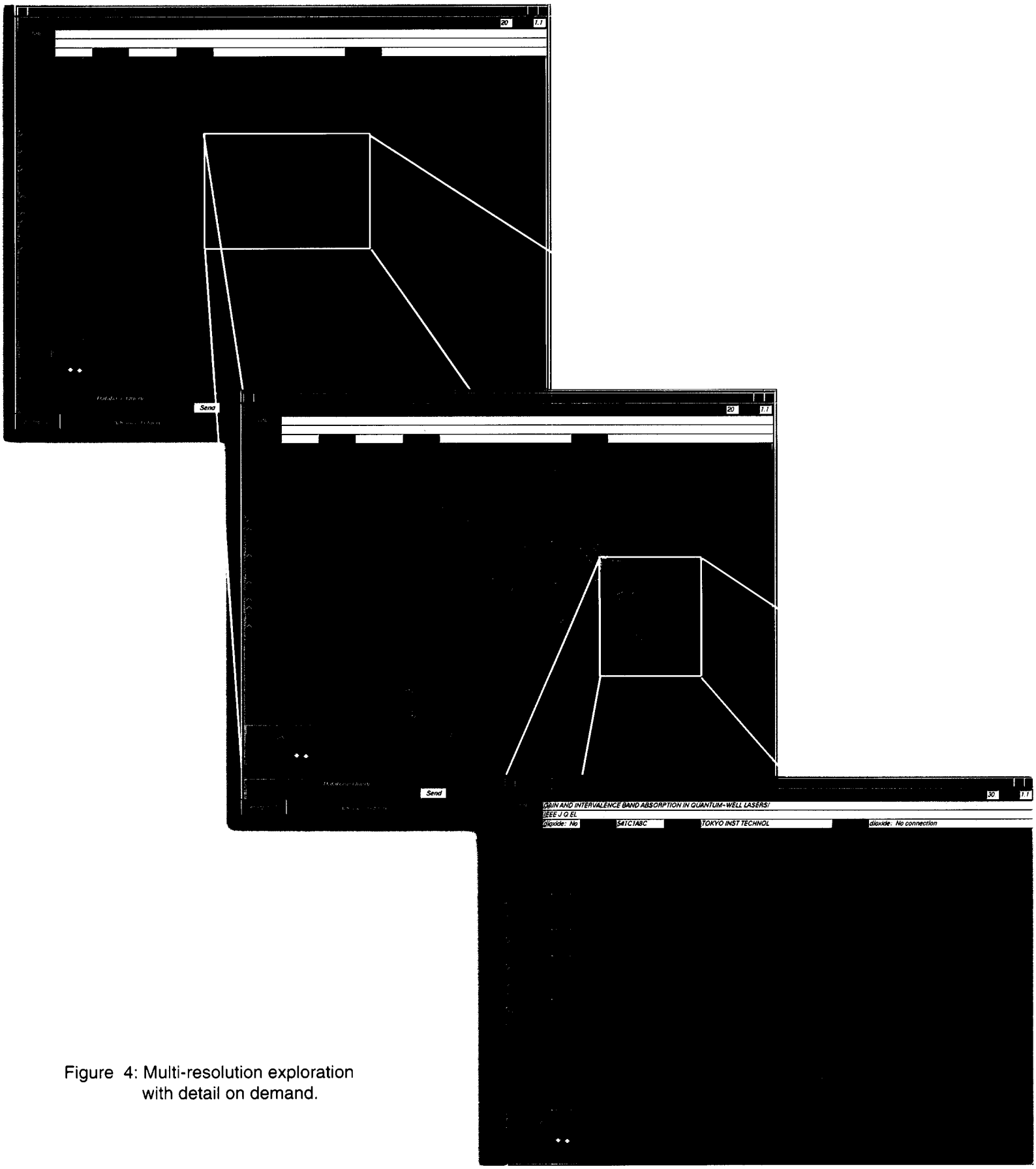

Figure 4: Multi-resolution exploration with detail on demand.

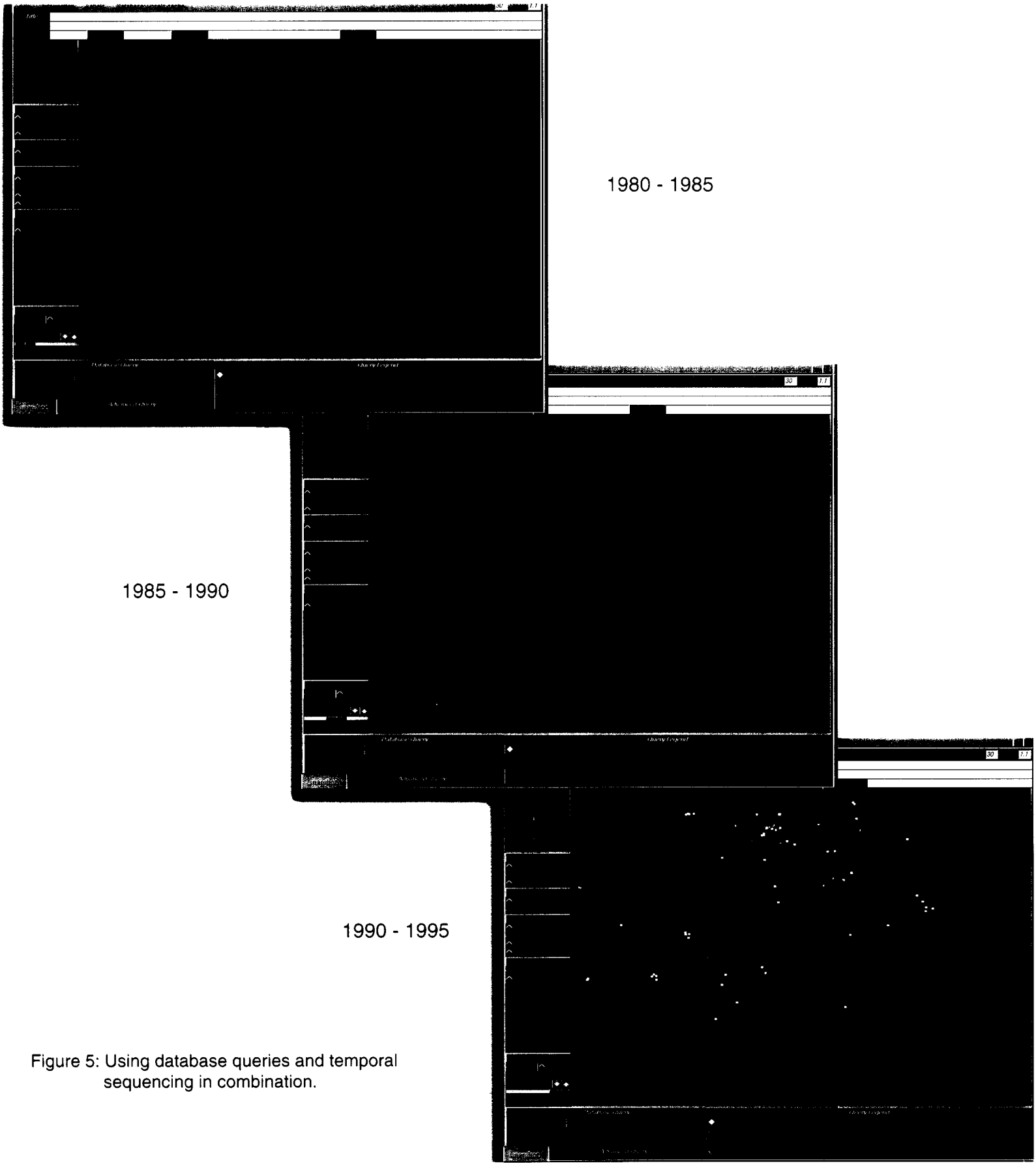

Figure 5: Using database queries and temporal sequencing in combination.

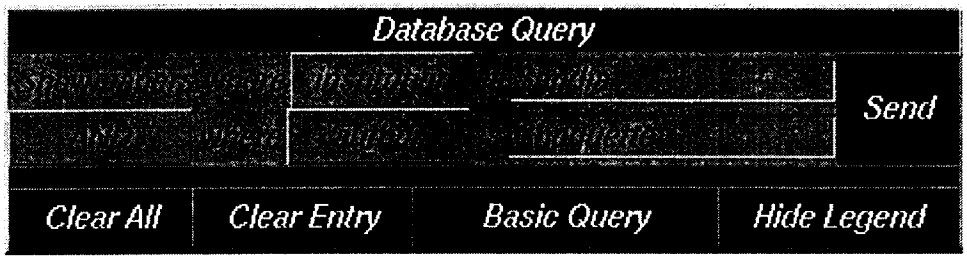

Figure 6. The SQL window where any of the database fields can be used to construct Boolean queries using conjunctions, disjunctions, and negation.